Semantic annotation of heritage building photos based on 3D spatial referencing

C. Busayarat, L. De Luca, P. Véron, M. Florenzano

Abstract — **Throughout this article, we present the principles of a semantic annotation using 3D model as a support to transfer the semantic layers to images. The approach could be used as an essential tool for the documentation of buildings of historical importance. Our research focused on the analysis and the implementation of tools and techniques for semantic annotation of photo, on its storage and retrieval from a database and on the manipulation of information in a real time 3D scene. Our approach, divided in three connected steps (semantic annotation of 3D model, spatial referencing of image and semantic annotation of image using 3D object's silhouette projection), has been finalized and tested on different archaeological sites. We have used these principles to develop an implementation model of a management and consultation system for data gathered from the Internet.**

Index Terms—**Semantic annotation, Spatial referencing, 3D, Historic architecture**

I. Introduction

In the domain of architectural survey, one of the most important data type is 2D visual information such as photo, drawing, painting or sketch including synthesis image created by computer or other recent technologies. When studying archaeological sites, the researcher generally collects a large amount of images. All of this information has been collected for research, references and will be used as data for future preservation or renovation in the future.

If one needs to recover specific information from a database, the efficiency of access, the rapidity of search and the capacity to sort the most relevant information are essential.

Nowadays, the most common method for architectural image searching uses the name of the object present in the image (e.g. capital, shaft, base, etc.). When the database is interrogated, a comparison is made between the input of the user and the semantic information pre-associated to images (the attributes). However, today, the method of association between semantic attributes and images is not entirely satisfactory. By lacking precision in detail, the image searching today still produces an inexact result or is missing the most accurate response.

C. B., L. D. L., M. F. Authors are with the UMR CNRS/MCC 694 MAP-GAMSAU, Marseille, France (corresponding author to provide phone: 33(0)491827170; fax: 33(0)491827171 ; e-mail: chawee.busayarat@ map.archi.fr, livio.deluca@ map.archi.fr, michel.florenzani@map.archi.fr).

P. V. Author is with the UMR CNRS 6168-LSIS, Aix-en- Provence, France, (corresponding author to provide phone: 33(0)442938124; e-mail: philippe.veron@aix.ensam.fr).

In this contribution we present our work and the methodology developed to assist the study of historical buildings and heritage sites. We are therefore proposing a new methodology that has high accuracy in details and precision in information association. This method is able to provide important information that reveals essential to the study and research in the archaeological field.

II. Approach overview

The semantic annotation of heritage and archaeological images has been receiving more and more attention during the recent years. A more practical way to classify, document and retrieve architectural information of historical site is a real need. Our objective is to explore and correctly use the 2D information for archaeological purposes. In this work, we developed a system to assist archaeologists in the digital classification, management and visualization of historical architecture images linked to existing databases.

Our approach of semantic annotation of images does not aim to create a direct relationship between image and semantics. Instead, we use 3D representation as a support between these two types of information.

Firstly, the 3D model is semantically annotated. The object semantics will be applied on to 3D model by breaking down the structures into their component parts (e.g. capital, shaft, base, etc.) following basic libraries of geometric primitives and associating each element to archaeological and architectural information extracted from existing databases.

Secondly, the relation between 3D model and 2D image is created by using spatial referencing system. This allows us to find the exact (or as close as possible) point of view of the image in 3D space. Different methods can be used in this task; depending on different image points of view, projection and image collecting process.

The last step is the silhouette projection from 3D model to spatial referenced images. As a result, we create semantic annotated vector images as a superimposed layer over the original image.

Because of the spatial referencing, these semantic annotated images will have the same point of view as the original images in the database. The final result of this method is a set of vector images that enclose the semantics of each architectural element in interactive way.

Our goal is also to create a web-based interface and system to retrieve and visualize 2D information based on the information produced by this method. This system can be used

by general public or serve professional needs such as those of architects and archaeologists.

III. Related works

A. Related works on semantic annotation of images

In many field of research, the semantic annotation of images is usually a creation of direct link between visual information and the meaning of the object in the image. We can store this relation into a database under the form of attributes. There are two types of methods to create this process.

The Manual method is the most simple and easiest to understand. Each image can be annotated by using different keywords (semantic) that show the meaning of the object in the image [1]. A connection between the images and key words must be manually preset beforehand. This process requires a manual manage by system administrator [2]-[3]. This method is more suitable for small and static system/database, because it requires a lot of information management done by a person. It is also difficult to update the system.

The Automatic method is a process of element detection in image using only computer algorithm. Normally, this process has two parts: image segmentation and object recognition.

The object class recognition can be achieved using a combination of particular models [4]-[6]. Many authors have considered these two tasks separately. For example [7] and [8] have considered only the segmentation task. Reference [9] shows image regions classification. Images are broken down into figures and background using a conditional random field model. Several authors such as [10] – [12] have considered recognition for multiple object classes. These techniques address image classification and object localization in fairly constrained images.

Our approach for semantic annotation does not aim to create a direct connection between image and semantics. Instead, we use 3D representation as a support between these two types of information. Therefore the semantic annotation of 3D models is also an important filed for our research.

B. Related works on semantic annotation of 3D model

The first example of 3D modeling and semantic classification was presented in [13]. Several researches concentrated on the development of classifications of architectural elements in theoretical frameworks [14] or in applications of the geometrical modeling [15].

Reference [16] presented a methodological approach to the semantic description of architectural elements based on theoretical reflections and research experiences.

In some research, the segmentation could be done automatically to analyze simple constrained object such as the floor or a step [17]. Others propose to process this task in a semi-automatic way, with more complex results [18]. Reference [19] also presented a system of data collection which allows creating semantic annotated 3D representation using open format and open application.

IV. Semantic annotation of 3D model

To construct a 3D representation of archaeological site, we employed Time of Flight laser scanners to survey and model some discoveries. The obtained point clouds were registered and meshed to create the 3D models. For some models additional surfaces have been introduced in order to rebuild hidden volumes and clearly separate the different architectural elements. Then our work continued with the segmentation phase and the linking to the database. This operation requires the help and support of archaeological and architectural knowledge to recognize transitions between different elements that constitute the architecture and manually segment it.

Furthermore, the identification of object leads to the semantic classification and object descriptions. The semantic segmentation is done directly on the 3D geometry managing by layers. Additional information such as position and orientation in 3D space are also added to the database. The semantic selecting of each single element depends on archaeological and architectural criteria [16].

I. Spatial referencing of photograph

Spatial referencing of images is a process that creates a relationship between 2D and 3D information. This process allows us to recreate a virtual camera of 2D image in 3D space. These cameras will imitate the behavior of cameras (in the case of photography) or the artist's point of view (in the case of painting, drawing or sketch) in the real world.

A. Image referencing methods

We have been able to obtain various types of image data (such as positional, directional and optical data) using two spatial referencing methods. Each image is suitable for different methods depending on the projection of the point of view and the processes of image creation and collection.

1) Manual method

This spatial referencing method is the most straightforward and is the easiest to implement. Various research aims to create a relationship between information presented in 2D and 3D, by approximately finding the point of view of an image by using other image, such as plan or elevation of building, as a support [2]. The position and orientation of a point of view can be represented in form of graphical symbol such as an arrow or a spot [3].

In case of using a 3D representation as a support [20], this method can use the virtual camera's image plane (textured by the image that needs to be referenced) as a reference to manually find the point and angle of view in the 3D model. This process takes more time than using a 2D support, but the result is much more precise and provides more spatial information.

To associate an image to its point of view, this method requires a manual intervention of the system administrator. However, human errors are possible and they reduce the precision of the system.

2) Semi-automatic method

The Semi-automatic spatial referencing method uses a programmed algorithm, called camera calibration, which

refers to the process of using numerical values to establish geometrical and optical parameters of a virtual camera which replicates the point of view of the image into the 3D model [21].

There are several methods to calibrate images. In this presentation we focused on Tsai algorithm [22]. When using this method, 3D points and 2D pixels need to be included to create the image. The Tsai method is using a two-stage calculation technique. First, the position and the orientation of the camera have to be calculated then the internal parameters are established.

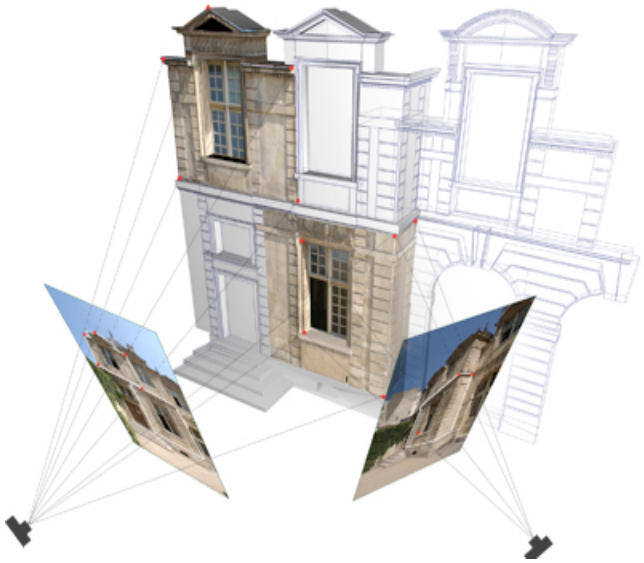

Fig.1: The calibration of the camera

The calibration method of the camera can be used to find the intrinsic parameters (focal distance, zoom and distortion), extrinsic parameters (matrix values with information about the rotation and translation of the camera), as well as the camera position in three-dimensional space.

B. The geometrical model of the camera

The result of the spatial referencing of an image is the geometrical model of the camera, which is a set of metric information that describes the characteristics of the camera in 3D space.

As there are many spatial reference methods, the development of our system had to solve the uncertainty related to the image positioning problem. Indeed, to be able to determine the position and orientation of an image in 3D, it was necessary to associate a geometrical model to the image.

Therefore two categories of geometrical information must be associated to each image.

- an external parameter: the relative position and orientation of the camera in a system of coordinates.
- an internal parameter: the camera's focal distance

The geometrical model of the camera in the 3D scene is expressed through the:

- T vector (Tx,Ty,Tz) – the position in space
- R vector (Rx,Ry,Rz) – the orientation in space
- Decimal digital value (FL) – the focal distance

The information is obtained from various spatial referencing methods and is stored in a database as attributes of the images. The images may then be stored in raster format and the 3D model stored in vector format. This set of information is essential for the next task of our work which is the silhouette projection process.

II. SEMANTIC ANNOTATION OF IMAGES

This task involves the projection of the silhouette of the 3D object onto a spatially referenced 2D image to create a semantic annotated image. The silhouettes are projected in the form of an additional vector layer of the original image (see Fig. 3). Each segment of a vector polygon corresponds to the shape of a 3D object in the scene as seen from the point of view of the image, and contains the same semantic attribute as the 3D element.

This process is entirely automatic and was developed using PHP (Hypertext Preprocessor) and MEL (Maya Embedded Language).The three types of information for this process are collected from the task beforehand: the original image, the camera's geometrical model (obtained from the spatial referencing) and the object's identity number and their semantic attributes (from semantic annotation of 3D model).

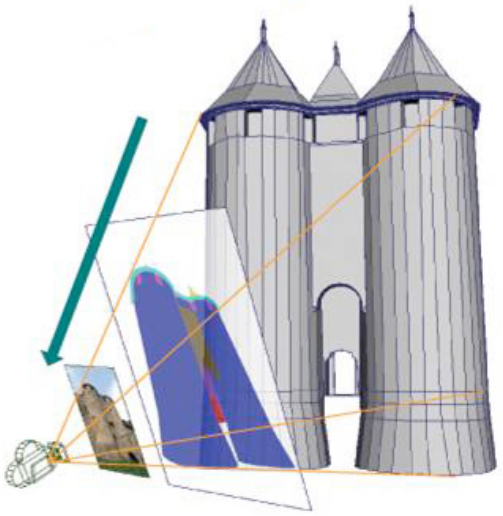

Fig. 3: silhouette projection onto spatial referenced image

To create the vector image from the 3D scene, we decided to use the vector rendering from a programmable 3D application (Maya). We developed a MEL script to automatically batch render cameras in the 3D scene. This script allows us to easily project the 3D object's silhouette onto the entire image in the database.

To communicate the data describing the geometrical model of the camera between the database and the 3D application, we use an editable text file as a support for data transfer. This file is created by a PHP page that is accessible only by the system administrator. The Maya script can read this external file and use its data to create virtual cameras in the 3D scene that mimic the behavior of the cameras in the real world.

After all the virtual cameras are created, the 3D application will retrieve the 3D objects from the object list into the database.

The batch rendering process produces a SVG (scalable

vector graphics) image file for each camera in the scene. The rendering is parameterized to calculate only the silhouettes of the object in the scene (including intersection edges) and ignore the others 3D polygon edges.

Each polygon segment is attached the object identity number and its semantic tag corresponds to its 3D counterpart. The commands in PHP page break down, restructure and edit the SVG image based on the information from archeological database.

The next process is an addition of interaction to the image using Javascript. The results of this process are interactive images that react to the user's mouse actions. The different polygons are colorized when the mouse rolls over them while text below the image also indicate the corresponded semantic of the active object. The tab below the image can be used for choosing between different renderings of the image such as colorized polygon, wire frame and raster image.

The area of each element present in the image is calculated using the position of the polygon vertices. This information can be useful in further development. For example, we can use this percentage as criteria for image searching or use for sorting selected information.

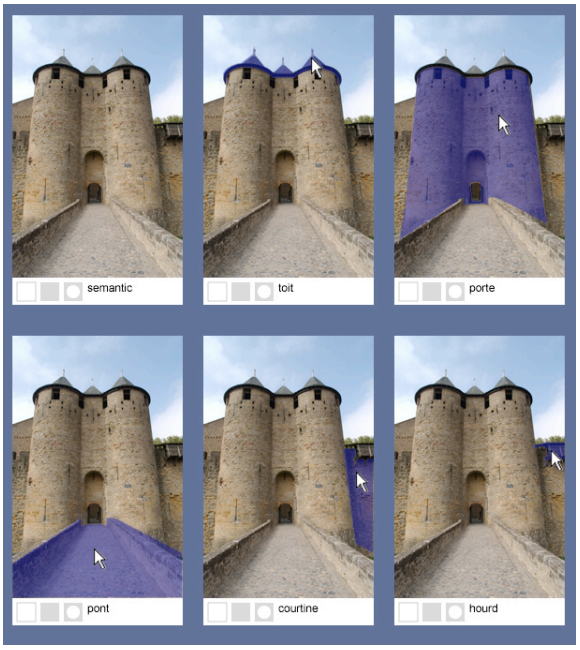

Fig. 4: interactive image displays different semantic of the object

The SVG interactive image is a hybrid representation that combines three different types of information: 2D image, 3D object and semantics. These representations allow us to view information in a new way. The information visualization using this method can be applied to systems designed to assist the archaeologists and the general public by allow them to understand the semantics of the architectural elements, their morphology, their textures etc.

V. INFORMATION VISUALIZATION

Our goal for this research is not only to create a semantic annotation system, but also to create a tool that allows us to take profit from the result of our method of semantic annotation. This tool allows users to navigate in real-time 3D and to observe a 3D model of the monument. At any moment, the system user can formulate image queries using the semantics of the object in the 3D scene as criteria.

After receiving the user's query, the system will produce a set of image that correspond the criteria. The user can then select an image that best corresponds to his need.

We are also suggesting another method to visualize the selected image information in the 3D space. This method allows the user to display the relation between 2D and 3D information in real-time interactive scene.

A. Image searching based on semantics

The image searching based on the semantics of an object is the main goal of our research. This function is a method of visualizing image information related to metric information in multi-media and the semantics of objects. It allows the user to select the semantics of architectural elements and launch requests to the database to find the images corresponding to the selected criteria.

For this function, the system automatically retrieves every available semantic in the database that corresponds to the selected project and creates a list for user to choose. One can also control the precision of image detection by giving the limit of the area representing the object in the image. Multiple selections can also be used. and the user can filter either including all the selected criteria or images that have at least one selected semantic ("and" and "or" in logic).

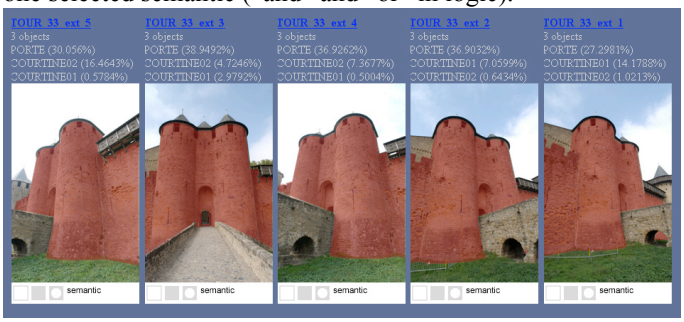

Fig. 5: result images using term "tower" and "Curtain wall" as criteria

After the query formulation, the system will find into the database the images corresponding to the selected criteria and copy each selected vector image (SVG) to a set of temporally image files. These temporally images will be automatically edited. The polygons that have semantic attribute corresponding to the selected semantic will be colored. This function allows the user to know the position of selected objects in the image. The resulting images are sorted according to the area of the object that has the corresponding attributes, then displayed in user interface

B. Searching an image point of view into the 3D scene

This function is used to find the position in 3D space of the point of view of a selected image. It allows the user to understand the relation between the selected image (2D information) and the 3D model.

The dialogue between the database and the 3D scene is used to search the point of view of the image. The camera data of

the selected image is then used to reproduce the parameters of the virtual camera in 3D space. The geometrical information concerning the camera (position, orientation and focal length) is transmitted through a program function which calculates the interpolation between them and the current position and the orientation of the navigation camera, as well as other values related to the selected image.

After the interpolation of the camera is finished, the system automatically displays a 2D plan in front of the navigation camera. The 2D plan is textured by the selected image and is posed on the 3D scene, exactly on the spot from where the real world photo was taken.

VI. CONCLUSION

Through our report we aim to design and develop a system of semantic annotation of images based on spatial referencing. In the same time we attempt to create a relationship between the three-dimensional information (3D model), the two-dimensional information (graphic documentary source) and the semantics of the architectural object.

This allows the evaluation at various levels of precision which we can obtain by superimposing visual 2D elements onto 3D scenes. The semantic annotation process is created by the projection of a semantic layer, based upon the silhouette of the 3D object, onto an image.

The level of precision of the correspondence between the semantically annotated polygon and original image depends on the accuracy of the spatial referencing process. The detail of semantic annotation is depending on the structure of the 3D decomposition and segmentation. However, the 3D decomposition can be redone or edited in case we desire to add more detail in the structure.

The system of information retrieval and visualization based on new information produced by our method of semantic annotation had been developed. Our system contains functionalities which can be used directly on the Web. This provides the possibility of a wide utilization of the system and makes the information accessibility possible for the general public as well as for professionals.

REFERENCES

[1] C. Chen, Documentation tools for management of cultural heritage for conservation of archeology, historical center, museums and archives: The case of global memory net, Journal of Zhejiang University, 2007.

[2] R. Kadobayashi, Y. Kawai, D. Kanjo, J. N. Yoshimoto, Integrated Presentation System for 3D Models and Image Database for Byzantine Ruins. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXIV, Part 5/W12, pp. 187192. 2003

[3] U. Herbig, P. Waldhäusl, APIS – Architecural Photogrammetry Information System.,1998.

[4] J. Winn, A. Criminisi, and T. Minka. Categorization by learned universal visual dictionary. In Proc. Int. Conf. on Computer Vision, volume 2, pages 1800–1807, Beijing, China, October 2005.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In Proc. IEEE Conf. Computer Vision and Pattern Recognition, volume 2, pages 264– 271, June 2003.

[6] A.C. Berg, T.L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In Proc. IEEE Conf. Computer Vision and Pattern Recognition, volume 1, pages 26–33, June 2005.

[7] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In Proc. Int. Conf. on Computer Vision, volume 2, pages 1150–1157, October 2003.

[8] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentations. In IEEE Workshop on Perceptual Organization in Computer Vision, volume 4, page 46, 2004.

[9] X. Ren, C. Fowlkes, and J. Malik. Figure. Ground assignment in natural images. In A. Leonardis, H. Bischof, and A. Pinz, editors, Proc. European Conf. on Computer Vision, volume 2, pages 614–627, Graz, Austria, May 2006. Springer.

[10] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. Proc. of CVPR 2004. Workshop on Generative-Model Based Vision., 2004.

[11] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 19(5):854– 869, May 2007.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006.

[13] P. Quintrand, J. Autran, M. Florenzano, M. Fregier, J. Zoller, La CAO en architecture. Hermes, Paris. 1985.

[14] A. Tzonis, L. Lefaivre. Classical Architecture - The Poetics of Order. The MIT Press, Cambridge. 1986.

[15] Gaiani, M. Translating the architecture of the world into virtual reality and vice-versa: 7 years of experimentation with conservation and representation at OFF. Proceedings of Heritage Applications of 3D Digital Imaging. Ottawa, Canada. 1999.

[16] L. De Luca., M. Florenzano, P. Veron. A generic formalism for the semantic modeling and representation of architectural elements. Visual Computer, 23, pp. 181-205. 2007.

[17] L. Iocchi, S. Pellegrini. Building 3D maps with semantic elements integrating 2D laser, stereo vision and INS on a mobile robot. In: 2nd ISPRS International Workshop 3D-ARCH., 2007.

[18] A. M. Manferdini1, F. Remondino, S. Baldissini1, M. Gaiani1, B. Benedetti, 3D modeling and semantic classification of archaeological finds for management and visualization in 3d archaeological databases, 2008.

[19] M. Lorenzini, Semantic approach to 3d historical reconstruction, Università degli Studi di Pisa, Pisa, Italy, 2009.

[20] C. Busayarat , L. De luca, P. Veron, M. Florenzano. An On-line system to Upload and Retrieve Architectural Documents Based on Spatial Referencing. Research in Interactive Design, vol. 3. Proceedings of IDMME-Virtual Concept 2008, Beijing, 2008

[21] R. Cipolla, T. Drummond. D. Robertson, Camera calibration from vanishing points in images of architectural scenes. Proc. 10th BMVC,pp. 382391., 1999.

[22] R. Y. Tsai , A versatile camera calibration technique for highaccuracy 3D Machine vision metrology using offtheshelf TV cameras and lenses ,IEEE Journal of Robotics and Automation, Vol. RA3, No. 4., 1987.